M.S. THESIS

Noise Robust Text-Dependent Speaker Verification Using Teacher-Student Learning Framework

교사-학생 학습 방법을 활용한 잡음에 강인한 화자 인식

BY

CHAE SEOK-WAN

August 2019

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

M.S. THESIS

Noise Robust Text-Dependent Speaker Verification Using Teacher-Student Learning Framework

교사-학생 학습 방법을 활용한 잡음에 강인한 화자 인식

BY

CHAE SEOK-WAN

August 2019

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Abstract

Noise robustness and speaker discrimination are two basic requirements for Speaker Verification (SV) systems to perform under noisy conditions. Noise robustness is the ability of the SV model to produce less deteriorated speaker embeddings in the presence of background noise. The speaker discriminative ability indicates how much the speaker embeddings can possess speaker-specific characteristics. A commonly used way to improve the noise robustness is to artificially add noises to the training data. However, this method has a disadvantage in that it weakens the speaker discriminability. This paper introduces a teacher-student learning framework for SV using parallel clean and noisy data to alleviate the known issue mentioned above. The baseline model, called "student network", is trained on both noisy speech and speaker embedding obtained from the model called "teacher network", which is trained on clean speech. In the scenario of using mobile devices, our text-dependent SV system based on self-attentive x-vector was evaluated on a keyword dataset. Experimental results show that the teacher-student framework is effective in alleviating the degradation of the model's speaker discriminative ability. This enables a decrease of the equal error rate under both clean and noisy conditions.

keywords: Noise Robust Speaker Verification, Teacher-Student Learning
student number: 2017-22872

Contents

List of Tables

List of Figures

Chapter 1

Introduction

화자 인증은 테스트 음성과 시스템에 등록된 음성으로부터 화자의 특성을 반영하는 정보를 추출하고 그 둘 간의 유사도를 계산하여 화자의 신원을 인증하는 과제이다. 최근에는 복잡하고 비선형적인 문제를 모델링하는 데 강점을 가진 심층신경망이 여러 분야에서 효과적으로 적용되고 있다. 이러한 심층신경망의 개략적인 발전과 함께 화자 인증 분야에서도 심층신경망을 이용한 연구들이 활발히 이루어지고 있다. d-vector[1], x-vector[2], end-to-end[6] 방법들이 그 예인데, 이들은 심층신경망을 이용해 추출한 화자 임베딩이 화자 간의 구분을 잘할 수 있음을 보여줬다. 그러나 이러한 연구들의 대부분은 배경잡음의 영향이 별로 없는 조용한 환경에서의 화자 인증 성능 만족을 우선적인 목표로 하였기 때문에, 실제로 많은 잡음이 혼입하는 실제 환경에서의 성능이 매우 저하될 수 있다. 따라서 화자 인증 모델이 잡음이 섞인 음성에 대해서도 덜 손상된 화자 임베딩을 추출할 수 있도록 하는 방법에 대한 연구가 여전히 해결해야 할 과제로 남아있다.

깨끗한 음성 데이터셋과 해당 음성에 대한 잡음이 섞인 데이터셋을 사용할 수 있다, 화자 인증 모델이 잡음에 강인해지도록 하는 방법에 대한 기존의 연구들은 크게 두 가지로 나누어서 볼 수 있다. 첫 번째로는 음성 향상 알고리즘을 화자 인증 모델의 전처리 알고리즘으로 사용하는 방법이다[7, 8]. 심층신경망을 이용하여 잡음이 섞인 음성으로부터 깨끗한 음성으로 mapping 해주는 음성 향상 모델을

만들고, 이 모델을 통해 향상된 음성을 화자 인증 모델에 사용하여 화자 인증이 잡음에 강인해지도록 하는 방법이다. 그러나 음성 향상 알고리즘은 음성 왜곡이 불가피하기 때문에 향상된 음성을 화자 인식에 사용하는 것은 최적의 솔루션이 아닐 수 있다. 또한 이 방법은 추가적인 전처리 모델이 필요하다는 단점이 있다. 두 번째로는 트레이닝 데이터에 다양한 잡음을 섞어서 모델을 학습시키고 이를 통해 모델이 잡음에 강인해지도록 하는 방법이다.[3, 6] 이 방법은 깨끗한 음성에 노이즈를 섞어줌으로써 트레이닝 데이터를 쉽게 늘릴 수 있으므로 효과적인 전략이라고 볼 수 있다. 그러나 이 방법은 모델이 노이즈에 강인해지는 대신 깨끗한 음성으로만 모델을 학습시켰을 때 보다 화자 구분 능력이 저하된다.

본 논문에서는 위에서 언급한 방법들 보다 효과적인 화자 인식 시스템 구성을 위해 교사-학생 학습 방법을 적용하였다. 기존의 교사-학생 학습 방법은 기계가 기계를 가르친다는 개념으로 intelligent teacher를 기계학습에 도입하여, 작은 복잡도 네트워크를 가진 학생 네트워크가 큰 복잡도를 가진 교사 네트워크의 솔루션을 모방할 수 있도록 하는 방법이다[9, 10, 11]. 음성 인식에서 모델 사이즈를 줄이는 방향으로 연구가 많이 진행되었으며, 몇몇 연구에서는 잡음에 강인한 음성 인식에도 응용이 되었다[12, 13, 14].

본 논문은 화자 인증 모델을 단순히 잡음이 섞인 음성으로 학습시키는 것이 아니라, 교사-학생 학습 방법을 화자 인증 시스템에 도입하여 성능을 향상시켰다. 교사 네트워크가 화자 구분을 잘할 수 있는 화자 임베딩을 출력할 수 있도록 깨끗한 음성 데이터셋으로만 학습시킨 후, 잡음이 섞인 음성으로 학생 네트워크를 학습시킬 때 교사 네트워크가 출력하는 화자 임베딩으로 부터 정보를 얻도록 한다. 이 방법은 잡음이 섞인 음성으로 학습시킬 때 얻게 되는 잡음에 강인함은 유지하면서도, 앞서 언급한 화자 구분 능력이 저하되는걸 완화할 수 있다. 본 논문이 제안하는 방법은 깨끗한 음성과 잡음이 섞인 음성 쌍을 이용하여 피처 간의 mapping을 해준다는 관점에서 음성 향상과 관련이 있다. 그러나 본 논문에서는 화자 인증에 더 적합하도록 향상 mapping을 front-end가 아닌 back-end에서 진행했다는 점에서 다르다.

본 논문에서 제안하는 방법에 대한 실험은 휴대 기기를 사용하는 시나리오상에

서 자체 녹음한 문장-종속 데이터셋으로 진행되었다. 데이터 셋은 200명의 화자가 ”Hi, Bixby”를 각각 100번씩 발화한 음성으로 구성되어 있으며, 화자 임베딩을 추출하는 네트워크로 self-attentive x-vector를 사용하였다.

본 논문의 구성은 하기와 같이 구성되었다.

먼저 2장에서는 심경심층망 기반의 화자 인증 시스템에 대한 간략한 설명과, 본 논문에서 사용한 x-vector 및 self-attentive x-vector에 대한 설명을 기술하였다. 3장에서는 교사-학생 학습 방법에 대한 간략한 설명과 본 논문에서 제안하는 방법인 교사-학생 학습 방법을 활용한 잡음에 강인한 화자 인식에 대하여 기술하였다. 4장에서는 실험 설정 및 실험 과정에 대하여 기술하였고, 5장에선 실험 결과와 분석 내용을 기술하였다. 그리고 마지막 6장에선 결론을 기술하였다.

Chapter 2

Deep Neural Network based Speaker Verification

2.1 System overview

심경심층망 기반의 화자 인증 시스템의 프레임워크는 그림 2.1 과 같다.

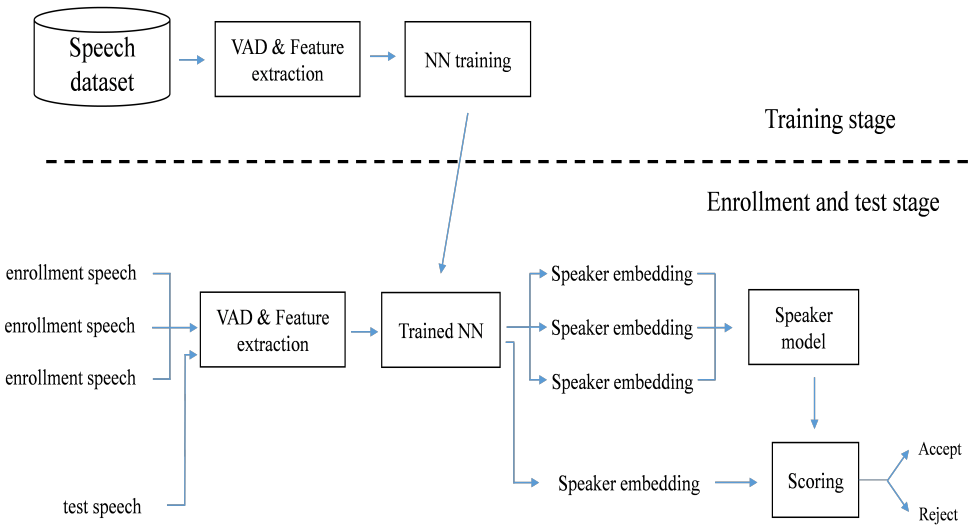

Figure 2.1: System overview.

2.1.1 Feature extraction

음성 신호를 네트워크의 입력으로 넣어 주기 전에 , VAD (Voice activity detection) 및 feature extraction 과정을 거친다. 음성 구간과 음성이 아닌 구간을 구분하는 것은 화자 인증하는 데에 있어서 중요한 부분이기 때문에, 화자 인증 시스템에서는 이 과정을 VAD 알고리즘을 통해 해결한다. 주로 사용되는 VAD 알고리즘으로는 에너지 기반 VAD 알고리즘이 많이 사용된다. VAD를 사용하여 음성 구간만을 선택한 후, 음성 특징 벡터를 추출하여 네트워크의 입력으로 넣어준다. 주로 사용되는 음성 특징 벡터로는 MFCC(Mel frequency cepstral coefficients)와 LPC(linear predictive coding) 등이 있다. MFCC는 인간의 청각 시스템의 특징을 반영하여 스펙트럼을 분석하고 음성의 특징을 추출하는 기법이다. 작은 크기 프레임 단위의 입력신호로부터 Fourier power spectrum을 계산한 후, 인간의 청각 시스템과 유사하게 설계된 Mel filter bank를 적용하고 각 필터의 에너지를 합하여 log를 취한다. 여기서 filter bank는 overlapping 되어 있기 때문에, filter bank 에너지들 사이의 상관관계를 줄여주는 DCT(Discrete cosine transform) 과정을 거쳐서 계산된다.

2.1.2 Speaker verification system

심경심층망 기반의 화자 인증 시스템의 프레임워크는 학습 단계와 등록 및 테스트 단계로 나눌 수 있다. 먼저 학습 단계에서 심경심층망 기반의 classification 네트워크를 구성한 후, 트레이닝 데이터셋 내의 화자를 구분할 수 있도록 cross entropy 손실 함수로 네트워크를 학습시켜준다. 학습이 완료되면 네트워크의 입력 음성에 대한 마지막 혹은 특정 hidden layer에서의 출력값은 화자를 구분할 수 있는 정보를 가지고 있다고 가정한다. 따라서 등록 및 테스트 과정에서는 출력 layer는 제거하고, 마지막 혹은 특정 hidden layer를 통해 입력 음성에 대한 화자 임베딩을 추출한다. 등록 단계에서는 시스템에 등록할 화자의 여러 음성을 학습된 네트워크에 입력으로 넣어주고, 네트워크가 출력하는 화자 임베딩들의 평균값으로 화자 모델을 만들어 시스템에 등록한다. 이후 테스트 문장을 학습된 네트워크의 입력으로 넣어서 얻은 화자 임베딩과 앞에서 시스템에 등록된 화자 모델과의 스코어를 계산한다.

스코어를 계산하는 방법으로는 코사인 유사도 혹은 PLDA(Probabilistic Linear Discriminant Analysis) 등이 있으며, 계산된 스코어가 문턱값(τ)을 넘으면 등록된 화자, 그렇지 않으면 등록되지 않은 화자로 판별한다.

2.1.3 Equal error rate

화자 인증에는 대표적으로 두 가지 에러가 있다. 등록되지 않은 화자의 음성으로부터 등록된 화자라고 판별하는 것(false accept, FA)과 등록된 화자의 음성으로부터 등록되지 않은 화자라고 판별하는 것(false reject, FR) 이다. 각각은 false-alarm과 miss error 라고도 불리며, 아래의 식과 같이 계산된다.

$$\text{False-Acceptance Rate (FAR)} = \frac{\text{Number of FA errors}}{\text{Number of imposter attempts}} \tag{2.1}$$

$$\text{False-Rejection Rate (FRR)} = \frac{\text{Number of FR errors}}{\text{Number of legitimate attempts}} \tag{2.2}$$

이 두 가지 에러는 문턱값에 의해서 달라지는데, 만약 문턱값이 너무 낮으면 FA error가 많아지고, 문턱값이 너무 높으면 FR error가 많아진다. 높은 보안이 요구되는 시스템에서는 문턱값을 높게 설정하여 FA error를 적게 나오도록 하고, 반대로 보안보다는 사용자의 편의성이 더 중요시되는 시스템에서는 문턱값을 낮게 설정하여 FR error 적게 나오도록 한다.

화자 인증 시스템을 다양한 동작 포인트에서 성능을 평가할 때 DET(detection error tradeoff) curve가 주로 사용한다. DET curve는 문턱값에 따른 FAR과 FRR을 그린 그래프로, 시스템의 성능이 더 좋을수록 curve는 원점에 가깝게 이동한다. 그림 2.2에서는 system 2가 모든 동작 포인트에서 원점에 더 가깝기 때문에 더 좋은 화자 인증 시스템이라고 할 수 있다. 그림 2.2과 같이 문턱값을 조절해 가면서 FAR과 FRR값이 같아지는 지점을 찾을 수 있고, 이때의 에러를 EER(Equal Error Rate)라고 하는데, 이는 화자 인증 시스템의 전반적인 성능을 평가하는 지표로 많이 사용된다.

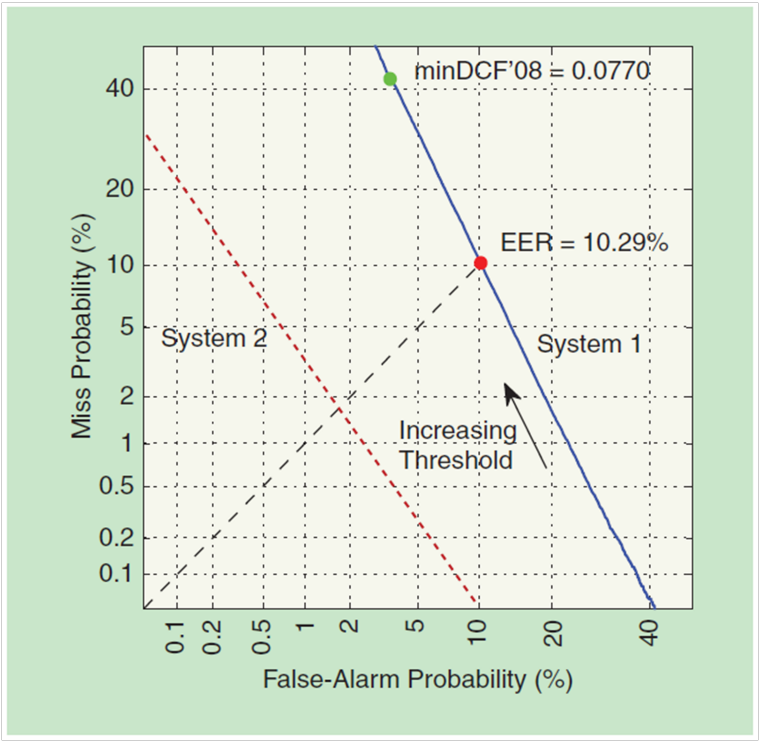

Figure 2.2: DET curves of two speaker verification systems[15].

2.2 Deep Neural Network Embeddings for Speaker Verification

2.2.1 X-vector

본 논문에서 채택된 심경신층망 기반 화자 인증 네트워크는 x-vector[2]로, 심층신경망을 이용하여 화자 구분이 가능 화자 임베딩을 출력하는 네트워크이다.

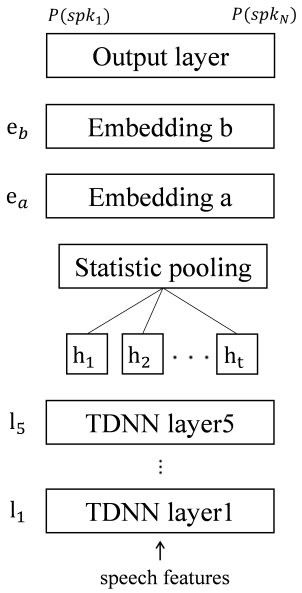

Figure 2.3: Sturcture of the DNN in the x-vector system.

네트워크의 구조는 그림 2.3 과 같다. 첫 번째부터 다섯번째 층까지는 TDNN(Time delay neural network)으로 구성되어있으며 프레임 단위로 동작한다. 만약 현재 time step이 t 라고 가정하면, 첫 번째 TDNN layer에서는 $\{t-2, t-1, t, t+1, t+2\}$ 의 시간적 맥락 정보를 접합하여 출력을 계산한다. 다음 두 번째 TDNN layer에서는 $\{t-2, t, t+2\}$ 의, 그리고 세 번째 TDNN layer에서는 $\{t-3, t, t+3\}$ 의 이전 layer 출력 값의 시간적 맥락 정보를 접합하여 출력한다. 네 번째와 다섯번째

TDNN layer에서는 시간적 맥락 정보를 다루지 않는다. 5개의 tdnn layer에서는 총 $(t-7)$ 부터 $(t+7)$ 까지의 시간적 맥락 정보를 다룬다.

Statistic pooling layer에서는 프레임 단위의 이전 레이어 출력값을 입력으로 받아서 문장 단위의 평균과 표준편차를 계산한다. 이렇게 계산된 문장 단위의 평균과 표준편차는 2개의 fully-coneected layer를 거친 후 마지막 softmax output layer를 통해 화자를 구분하게 된다. 이 두 개의 fully-connected layer는 화자를 구분할 수 있는 정보를 담고 있다는 가정하에, 이 layer들의 출력값을 화자 임베딩으로 사용된다. 이전 d-vector 방법은 네트워크가 프레임 단위에서 화자 임베딩을 출력하고 그 값을 평균 내어 문장 단위의 화자 임베딩을 계산하는 반면, x-vector에서는 statistic pooling layer를 통해 네트워크 중간에서 프레임 단위의 결과를 문장 단위의 결과로 바꿔주고 최종적으로 문장 단위의 화자 임베딩을 출력한다.

본 논문에서는 마지막 hidden layer에 L2 - normalization 을 적용하였다[5]. 마지막 hidden layer의 입력 e로 부터 출력 c는 아래의 식 2.3 과 같이 계산된다.

$$c = w * \frac{e}{\|e\|_2}, \left(\|e\|_2 = \sqrt{\sum_i |e_i|^2} \right), \tag{2.3}$$

여기서 w는 학습되는 파라미터이며, 출력되는 화자 임베딩의 각 성분 크기가 너무 작아지지 않도록 하여 학습 속도가 저하되는 것을 방지해준다.

2.2.2 Self-attentive x-vector

본 논문에서는 잡음 환경에 더 적합한 기존 x-vector에 self-attention 메카니즘을 적용한 모델[4]을 채택하였다. 기존의 x-vector 시스템은 앞서 언급했던 것처럼 statistic pooling layer에서 프레임 단위 출력의 평균과 표준편차를 계산해주는데, 이는 모든 프레임 단위의 출력을 동등하게 계산해주는 것과 같다. 그러나 모든 프레임이 동등하게 중요한 정보를 가지고 있지 않기 때문에, 이를 개선하기 위해 attention 메카니즘을 도입한 새로운 x-vector 모델이 제안되었다.

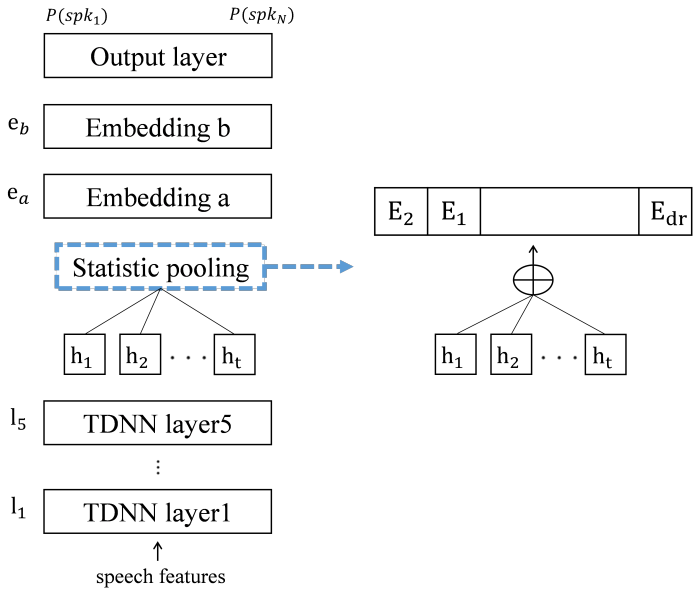

Figure 2.4: Structure of the self-attention layer.

그림 2.4 에 묘사된 것과 같이 기존의 statistic pooling layer를 self-attention layer로 대체하여, 이전 hidden layer에서 프레임 단위로 출력되는 값의 가중 평균과 가중 표준편차를 계산한다. 여기서 가중치는 화자를 분류하는 성능이 최대가 될 수 있도록 attention 메카니즘을 통해 트레이닝 중에 학습된다. attention 메카니즘은 다음과 같다.

만약 attention layer 아래 l_5 layer의 출력이 $H = \{h_1, h_2, h_3, ..., h_T\}$ 라고 가정

하자. attention 메카니즘은 이 H 를 입력으로 받아서 가중 행렬 A를 다음과 같이 계산한다.

$$A = g_1(g_2(H^T W_1)W_2)). \tag{2.4}$$

여기서 $g_2(\cdot)$ 는 ReLU 활성 함수이고, $g_1(\cdot)$ 는 열-단위로 계산되는 softmax 활성 함수로 가중치의 값이 0에서 1사이 값이 되고, 그 값들의 합이 1이 되도록 해준. 이렇게 계산된 A의 각 열벡터는 h_t에 대한 가중치를 나타낸다. 이제 가중 평균은

$$E = HA \tag{2.5}$$

로 계산된다. 만약 h_t 의 차원이 d_h 라면 H, W_1 그리고 W_2의 차원은 각각 $d_h \times T$, $d_h \times d_a$, 그리고 $d_a \times d_r$이 된다. d_r을 증가시켜 Multi-head attention을 쉽게 설정할 수 있는데, 이는 다양한 관점에서 음성의 정보를 확인할 수 있게 해준다. Multi-head attention ($d_r > 1$)을 사용할 때, 가중치 행렬 A가 다양하게 학습되게 하기 위해 다음과 같은 패널티 함수를 설정해준다.

$$P = \left\|(A^T A - I)\right\|_F^2 \tag{2.6}$$

여기서 I 는 identity matrix 이고 $\|\cdot\|_F$ 는 Frobenius norm이다. P는 기존 손실 함수와 함께 사용한다.

전처리로 사용하는 VAD 알고리즘은 잡음 환경에서 음성 구간을 찾는데 문제가 있을 수 있다. 따라서 본 논문에서는 VAD 과정을 attention 메카니즘으로 대체하여 네트워크가 화자 인증에 필요한 음성 구간을 학습을 통해 찾도록 하였다.

Chapter 3

Proposed method for noise robust speaker verification

3.1 Teacher-student learning framework

일반적으로 학습 데이터의 양이 많거나, 고차원 피처를 다루는 문제들은 심경 심층망의 구조가 크고 깊을수록 더 좋은 성능을 보인다. 또한 여러 구조의 심경 심층망 구조를 만든 뒤에 이들로부터 나오는 결과를 종합하는 앙상블 방법 또한 성능을 올리는 방법으로 많이 사용된다. 그러나 이러한 크고 깊은 모델이나 앙상블 된 모델은 계산량이 많아서 backward와 forward가 오래 걸린다는 단점이 있다. 따라서 크고 복잡한 모델의 구조를 작게 만드는 방법들이 다양하게 연구되고 있는데, 교사-학생 학습 방이 그중 하나다.

교사-학생 학습 방법은 큰 복잡도를 가진 교사 네트워크와 비교적 작은 복잡도를 가진 학생 네트워크로 구성되어있다. 성능이 좋은 교사 네트워크를 먼저 학습시킨 다음, 학생 네트워크를 학습할 때 교사 네트워크로부터 가이드를 받게 하여, 학생 네트워크가 교사 네트워크를 모방할 수 있도록 한다. 교사-학생 학습 방법을 knowledge distillation이라고도 부르는데, 이 방법을 통해 오로지 학생 네트워크로만 학습했을 때보다 성능이 잘 나오도록 할 수 있다.

Knowledge distillation 방법으로는 softmax output layer의 출력인 class 확률값을 transfer 하는 방법[10]과 hidden layer의 출력값을 transfer 하는 방법[11]등이 있다. 먼저 softmax output layer의 출력인 class 확률값을 transfer 하는 방법은 다음 그림 3.1 과 같다.

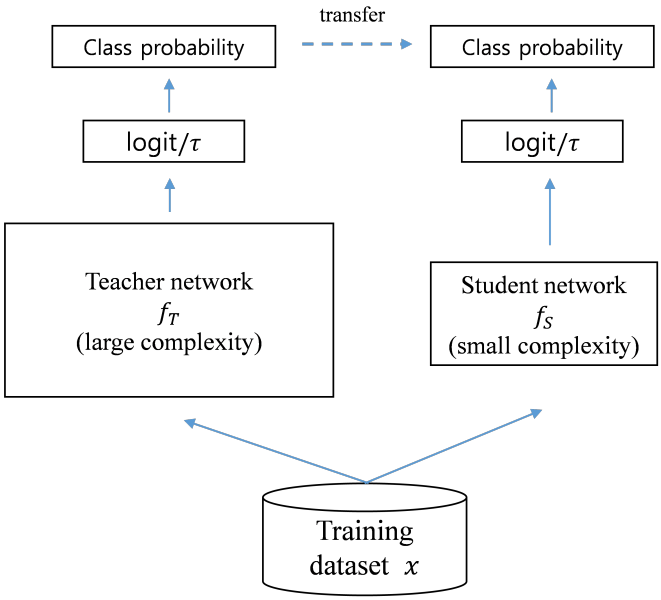

Figure 3.1: Distilling the knowledge through the softmax output layer.

교사 네트워크와 학생 네트워크의 logit을 temperature τ 로 normalize 한 후, normalized된 logit의 class probabilty를 transfer 하는 방식으로 아래의 식 3.1 과 같다.

$$L_{KD} = KL(softmax(\frac{f_T(x)}{\tau}), softmax(\frac{f_S(x)}{\tau})), \tag{3.1}$$

여기서 $KL(\cdot)$는 Kullback–Leibler divergence 이며, 기존의 cross entropy 손실 함수와 함께 사용된다.

Hidden layer의 출력값을 transfer 하는 방법은 아래의 그림 3.2 와 같다.

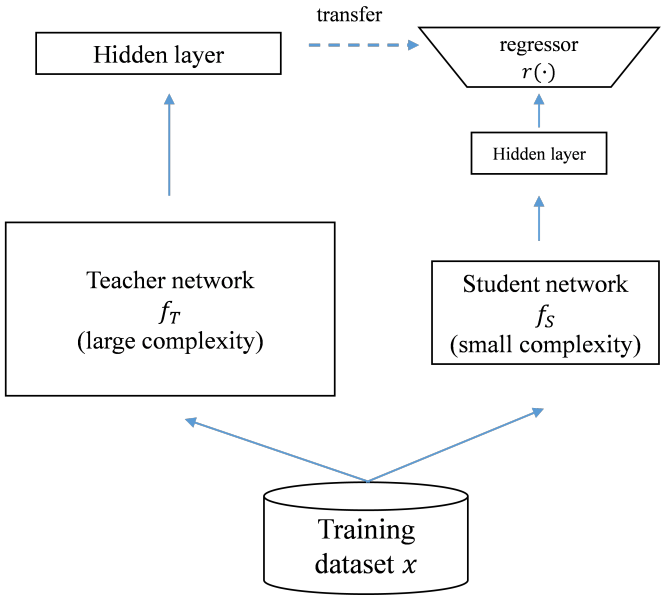

Figure 3.2: Distilling the knowledge through the hidden layer.

교사 모델과 학생 모델의 hidden layer의 크기가 다를 수 있기 때문에, 아래의 식 3.2 와 같이 regressor 함수를 사용하여 두 hidden layer 출력의 크기를 맞춰준 다음 transfer 되도록 한다.

$$L_{HT} = \|f_T(x) - r(f_S(x))\|_2^2. \tag{3.2}$$

3.2 Teacher-student learning for noise robust speaker verification

화자 인증 시스템은 잡음 환경에서 기본적으로 두 가지가 요구된다. 첫 번째는 잡음에 강인함으로, 화자 인증 시스템이 잡음 환경에서도 덜 손상된 화자 임베딩을 만들어내는 것이다. 두 번째는 화자 구분 능력으로, 화자 인증 시스템이 만든 화자 임베딩이 화자 간의 구분이 가능한 화자 특성을 많이 포함하는 것이다. 화자 인식 시스템이 잡음에 강인해지도록 하기 위한 방법으로 잡음이 섞인 음성 데이터로 화자 인식 모델을 학습시키는 방법을 많이 사용한다. 이 방법으로 모델이 잡음에 강인해질 수 있지만, 깨끗한 음성 데이터만으로 모델을 학습시켰을 때 보다 화자 구분 능력이 저하될 수 있다.

본 논문에서는 이를 개선하기 위해 단순히 잡음이 섞인 음성 데이터로 모델을 학습시키는 것이 아니라, 앞 장에서 언급한 교사-학생 학습 방법을 화자 인식 시스템에 적용해서 화자 구분 능력이 저하되는 것을 완화할 수 있게 하였다. 먼저 깨끗한 음성 데이터만으로 교사 네트워크를 미리 학습시켜서, 교사 네트워크가 출력하는 화자 임베딩이 잡음이 섞인 음성 데이터로 학습되는 모델이 출력하는 화자 임베딩 보다 화자 간의 구분을 잘할 수 있는 특징이 많이 포함할 수 있도록 해준다. 학생 네트워크는 잡음이 섞인 음성 데이터로 학습을 하게 되는데, 이때 미리 학습된 교사 네트워크가 출력하는 화자 임베딩으로 부터 참조하면서 학습되도록 한다. 교사 네트워크는 학생 네트워크를 가이드 하는 역할로만 쓰이기 때문에, 학생 네트워크를 학습할 때 교사 네트워크의 파라미터는 업데이트되지 않도록 하였다. 앞 장에서 설명한 기존의 교사-학생 학습 방법과는 다르게 본 논문에서는 학생 네트워크가 더 큰 모델로 부터 배우는 것이 아니라 다른 환경으로 학습된 모델로 부터 배우도록 구성하였다. 따라서 교사 네트워크와 학생 네트워크의 구조는 같게 설정 하였다.

3.2.1 Initialization scheme

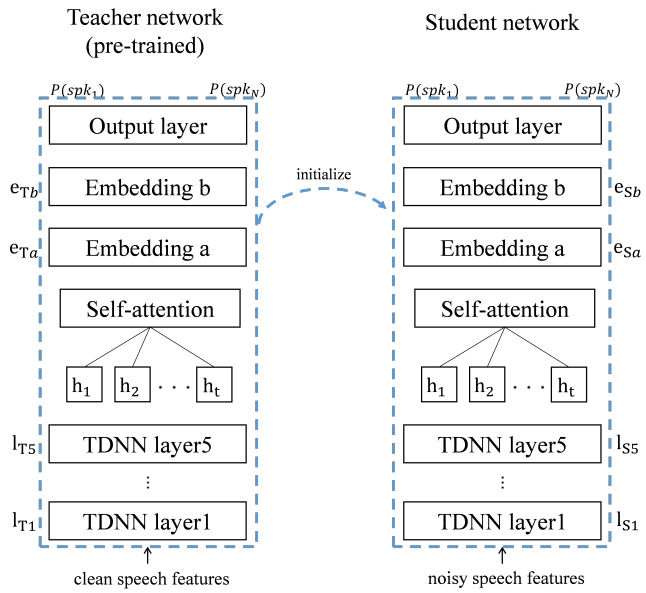

Figure 3.3: Proposed method for robust speaker verification : teacher initialization scheme.

학생 네트워크를 학습할 때 네트워크의 파라미터 초깃값으로 두 가지가 사용될 수 있다. 첫 번째로는 랜덤한 값으로 학생 네트워크를 초기화할 수 있다. 두 번째로, 교사 네트워크와 학생 네트워크의 구조가 같다면, 미리 학습된 교사 네트워크의 파라미터들을 사용할 수 있다. Darrell의 연구[16] 결과에서 미리 학습된 교사 네트워크의 파라미터를 사용하는 것이 더 좋은 성능을 보였다. 본 논문에서도 그림 3.3 과같이 미리 학습된 교사 네트워크의 파라미터를 학생 네트워크의 파라미터 초깃값으로 사용하였다. 이는 attention 메카니즘으로 VAD 대신 음성 구간을 찾을 때, 깨끗한 음성 데이터로만 학습한 모델이 잡음이 섞인 음성 데이터로 학습된 모델보다 인식에 필요한 음성 구간을 더 잘 찾을 것이기 때문에 랜덤 값으로 학생 네트워크를 초기화하는 것보다 더 좋은 효과를 기대할 수 있다.

3.2.2 Objective function

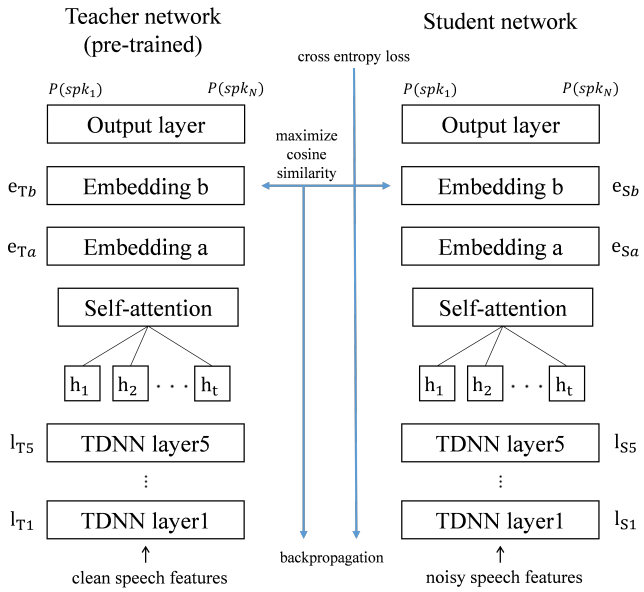

Figure 3.4: Proposed method for robust speaker verification : objective function.

본 논문에서 학생 네트워크는 두 가지 목적 함수에 의해 학습된다. 첫 번째는 기본적으로 화자 인증 네트워크를 학습할 때 사용되는 cross-entropy 손실 함수이다.

$$L_{CE} = -\sum_{r} S_r log P(spk_r|x_{1:T}), \tag{3.3}$$

여기서 S_r 은 트레이닝 라벨로써 입력 $x_{1:T}$ 이 화자 spk_r 의 발화이면 1 이고, 다른 화자의 발화이면 0 이다.

본 논문에서 제안하는 방법의 목표는 학생 네트워크가 미리 학습된 교사 네트워크의 화자 구분 능력을 모방하는 것이다. 그림 3.4 와 같이 따라서 잡음이 섞인 음성 데이터로 학생 네트워크를 학습할 때, 깨끗한 형태의 동일한 음성 데이터를 미리 학습된 교사 네트워크의 입력으로 넣어서 나오는 embedding layer의 출력값과 학생 네트워크의 embedding layer의 출력값의 거리를 줄이는 방향으로 학습시켜

서, 교사 네트워크로부터 가이드를 받을 수 있도록 해준다. 본 논문에서는 마지막 hidden layer의 출력값만 화자 임베딩으로 사용하였기 때문에 Embedding b layer를 통해 가이드를 받도록 설정하였고, 출력값 사이의 거리를 아래의 식 3.4 와 같이 코사인 유사도로 계산하였다.

$$L_{TS} = -\frac{(e_{Tb})^T(e_{Sb})}{\|(e_{Tb})\| \, \|(e_{Sb})\|}, \tag{3.4}$$

여기서 코사인 유사도를 최대화하도록 목적 함수를 설정하였기 때문에, 식 앞에 음의 부호가 붙는다. 기본적인 cross entropy 손실 함수 3.4 와 제안된 목적함수를 imitation parameter γ로 아래의 식 3.5과 같이 결합하여 학생 모델을 학습시킨다.

$$L = (1-\gamma)L_{CE} + \gamma L_{TS}. \tag{3.5}$$

$\gamma = 0$ 이면 전체 목적 함수가 기본적인 화자 인식에서 사용하는 목적 함수와 같지만, 교사 네트워크 파라미터 값을 학생 네트워크 파라미터의 초깃값으로 사용했다는 점에서 베이스 라인과 다르다. $\gamma > 0$ 로 증가시키면서 본 논문에서 제안하는 목적 함수의 비중을 늘릴 수 있다.

Chapter 4

Experimental setup

본 논문에서 제안하는 방법인 교사-학생 학습 방법으로 학습한 학생 네트워크, 베이스라인인 교사 네트워크를 사용하지 않고 학습한 학생 네트워크 그리고 교사 네트워크로 구성하여 실험하였다.

미리 학습된 교사 네트워크의 파라미터값으로 학생 네트워크를 초기화시킨 경우를 제외한 나머지 네트워크들은 Xavier initialization[19]을 사용하여 초기화 시켰다. Tensorflow toolkit[17]을 사용하여 실험하였으며, 모델 최적화 방법으로 Adam optimizer[18]를 사용하였다.

4.1 Model structure

음성 입력에 12.8ms 길이로 6.4ms씩 shift 되는 윈도우를 씌워서 20-차원의 MFCC를 계산하였다. 델타와 델타-델타값을 사용하여 총 60-차원이 되도록 구성한 후 입력 feature로 사용하였다. 전처리 과정에서 사용되는 VAD 알고리즘은 잡음 환경에서 문제가 있을 수 있기 때문에 attention 메카니즘으로 대체하였다.

교사 네트워크와 학생 네트워크의 구조는 표 4.1 에 기술하였으며, 두 네트워크는 동일한 구조로 설정하였다. 첫 번째 layer l_1 부터 l_5까지는 TDNN later로 구성하였다. 현재 time step이 t 라고 가정하면, l_1 layer에서는 $(t-2)$ 에서 $(t+2)$

까지의 시간적 맥락 정보를 접합하여 출력을 계산하고, l_2 layer에서는 $\{t-2, t, t+2\}$ 만큼의, l_3 layer에서는 각각 $\{t-3, t, t+3\}$의 이전 layer 출력 값의 시간적 맥락 정보를 접합하여 출력을 계산한다. l_4 와 l_5 layer에서는 시간적 맥락 정보를 다루지 않고, 5개의 TDNN layer에서 총 $(t-7)$ 부터 $(t+7)$ 까지의 시간적 맥락 정보를 다룬다. l_1 layer는 1536개의 노드를, l_2 와 l_3 layer는 512개의 노드를, l_4 와 l_5 layer는 256개의 노드를 가진다. self-attention 메카니즘에서 d_a는 128로, d_r은 5로 설정하였다. embedding layer a(e_a) 와 embedding layer b(e_b)는 fully-connected layer로 구성하였으며, 각각은 500개와 300개의 노드를 가진다. l_1 에서 l_5 까지의 TDNN layer들은 ReLU 활성 함수를, 출력 layer는 softmax 활성 함수를, e_a 와 e_b layer는 linear를 사용하였다. 본 논문에서는 네트워크의 e_b layer의 출력값을 화자 임베딩으로 사용하였고, 등록 및 테스트 과정에서는 코사인 유사도를 사용하여 스코어를 계산하였다.

Layer	**Layer context**	**# of node**	**Activation**
TDNN layer1(l_1)	$[t-2, t+2]$	1536	ReLU
TDNN layer2(l_2)	$\{t-2, t, t+2\}$	512	ReLU
TDNN layer3(l_3)	$\{t-3, t, t+3\}$	512	ReLU
TDNN layer4(l_4)	$\{t\}$	256	ReLU
TDNN layer5(l_5)	$\{t\}$	256	ReLU
self-attention	$d_a = 128, d_r = 5$, g_1=sofmax, g_2 = ReLU		
embedding a(e_a)	-	500	Linear
embedding b(e_b)	-	300	Linear
output layer	-	N	Sofmax

Table 4.1: The embedding DNN architecture.

4.2 Dataset

본 논문에서 제안하는 방법에 대한 실험은 휴대 기기를 사용하는 시나리오상에서 자체 녹음한 문장-종속 데이터셋으로 진행되었다. 전체 데이터셋은 깨끗한 환경에서 남자 화자 93명, 여자 화자 107명으로 구성된 200명의 화자가 "Hi, Bixby"를 100번씩 발화한 음성으로 구성되어 있으며, 음성 데이터의 sampling rate는 20kHz이다. 그중 남자 화자 75명, 여자 화자 85명으로 구성된 총 160명의 발화를 선택하여 트레이닝 데이터셋으로 사용하였고, 나머지 40명의 발화는 등록 및 테스트 데이터셋으로 사용하였다.

ITU-T recommendation P.501[20]의 {Cafeteria, car_mono1, con_mono1, INCAR, Kids1, met_mono1, rai_mono1, Traffic1, off_mono1, res_mono1_30s, STREET}의 노이즈 데이터와 NOISEX-92[21]의 {factory1,volvo, white16k, babble}의 노이즈 데이터를 앞에서 언급한 깨끗한 음성 데이터에 인위적으로 더해서 잡음이 섞인 음성 데이터를 만들었다. 각각의 노이즈는 20kHz로 re-sampling 하였다.

4.2.1 Training data

본 논문에서 실험에 사용한 트레이닝 데이터셋은 160명이 100번 발화한 160 × 100개의 깨끗한 음성 데이터셋과 동일한 음성에 대한 잡음이 섞인 버전의 160 × 100개의 음성 데이터셋이다. 잡음이 섞인 음성 데이터셋은 깨끗한 음성 데이터셋을 대략 11×8개의 subset으로 나눈다음, 각각의 subset에 11개의 노이즈{Cafeteria, car_mono1, con_mono1, factory1, INCAR, Kids1, met_mono1, rai_mono1, Traffic1, volvo, white16k} 를 다양한 signal-to-noise ratio(SNR) {-5, 0, 5, 10, 15, 20, 30 and clean condition}로 균등하게 더해서 구성하였다.

4.2.2 Enrollment and test data

등록 및 테스트 과정에서는 트레이닝에 사용된 160명을 제외한 40명의 화자의 발화가 사용되었다. 등록 과정에서는 각 화자당 4문장이 사용되었으며, 테스트 과정에서는 나머지 80문장이 사용되었다. 각 화자당 80번의 target trial과 80×39

번의 non-target trial을 계산하였다. 본 논문에서는 사용자가 모바일 기기 시스템에 화자를 등록하는 과정은 깨끗한 환경에서 진행한다고 가정하였다. 따라서 등록 과정에서는 깨끗한 음성 데이터셋을 사용하였다. 테스트 데이터셋은 깨끗한 음성 데이터셋과 잡음이 섞인 음성 데이터셋 두 가지를 사용하였다. 잡음이 섞인 음성 데이터셋은 깨끗한 음성 데이터셋에 모바일 기기를 사용할 때 흔히 발생할 수 있는 노이즈{babble, off_mono1, res_mono1_30s, STREET}를 선정하여, SNR {0, 5, 10, 15 and clean condition}로 균등하게 더해서 구성하였다.

'Hi, Bixby'	**Training set**	**Enroll & test set**
Speakers	160 spekaers (75 male, 85 female)	4 spekaers (18 male, 22 female)
Noise type	{Cafeteria, car_mono1, con_mono1, factory1, INCAR, Kids1, met_mono1, rai_mono1, Traffic1, volvo, white16k}	{babble, res_mono1 , off_mono1, STREET} res_mono1
SNR	{-5, 0, 5, 10, 15, 20, 30dB and clean}	{-5, 0, 5, 10, 15dB and clean}
Utterances	100	80 target trials 80×39 non-target trials

Table 4.2: Dataset.

Chapter 5

Results

본 5장의 실험 결과에서 proposed는 본 논문에서 제안하는 교사-학생 학습 방법으로 학생 네트워크를 학습시킨 결과이다. proposed 1은 식 3.5 에서 γ를 0으로 설정하였고, proposed 3은 식 3.5 에서 γ를 1로 설정하였다. proposed 2에서는 γ를 0.1로 설정하여, 두 개의 비용 함수 L_{CE}와 L_{TS}의 dynamic range가 비슷해지도록 하였다.

실험 방법에 따른 학습 데이터셋의 구성은 아래의 표 5.1과 같다.

Training method	**training data set**
Clean training (teacher only)	160 × 100 clean
Multi-style training 1 (student only)	160 × 100 noisy
Multi-style training 2 (student only)	160 × 100 clean & 160 × 100 noisy
Proposed 1, 2, 3 (teacher-student)	160 × 100 clean & 160 × 100 noisy

Table 5.1: Training data set used in each method.

Training method	clean test set	noisy test set
Clean training (teacher only)	1.55	13.93
Multi-style training 1 (student only)	2.84	3.61
Proposed 1 (teacher-student)	2.38	3.49
Proposed 2 (teacher-student)	2.19	3.35
Proposed 3 (teacher-student)	1.95	2.99
Multi-style training 2 (student only)	2.27	3.48

Table 5.2: Comparison of clean training, multi-style training and proposed method with different imitation parameters. Results are reported in equal error rate (EER) [%].

5.1 Clean training versus multi-style training

표 5.2 의 clean training과 multi-style training 1의 결과를 비교해보면, 트레이닝 데이터셋에 다양한 잡음을 섞어줌으로써, 잡음 테스트셋에 대 성능이 상당히 많이 오른 것을 확인할 수 있다. 그러나 깨끗한 음성 테스트셋에 대해서는 반대로 성능이 저하되는 것을 확인할 수 있는데, 이는 화자 구분 능력이 저하된 것이라 볼 수 있다. multi-style training 2 결과를 보면 깨끗한 음성 데이터셋과 해당 음성에 잡음을 추가하여 만든 잡음이 섞인 음성 데이터셋 모두를 트레이닝 데이터로 사용하였을 때, 깨끗한 환경과 잡음 환경 모두에서 multi-style training 1보다 좋은 결과를 얻을 수 있었다. 그러나 clean training에서 사용한 깨끗한 음성 데이터셋을 모두 사용하였음에도 불구하고, 여전히 화자 구분 능력이 저하된 결과를 얻을 수 있었다. 본 연구는 위와 같이 잡음이 섞인 음성으로 네트워크를 학습 시켜줄 때 화자 구분 능력이 저하되는 것을 완화하기 위해 시작되었다.

5.2 Initialization scheme

표 5.2 에서 proposed 1 은 식 3.5 에서 imitation parameter γ 가 0으로 설정되었기 때문에 multi-style training 1과 동일한 목적 함수 (식 3.3) 로 학습되었다. 학생 네트워크 파라미터의 초깃값으로, proposed 1은 미리 학습된 교사 네트워크의 파라미터를 사용하였고, multi-style training 1은 랜덤값을 사용하였다. 이 결과로부터 미리 학습된 교사 네트워크의 파라미터를 쓰는 것이 더 효과적인 것을 확인할 수 있었다. 따라서 본 논문이 제안하는 방법인 proposed 1, 2, 3은 이와 같은 방법으로 초기화시켰다.

5.3 Baseline versus proposed method

표 5.2 에서 imitation parameter γ 가 증가할수록 (proposed 1 에서 3), 성능이 오르는 것을 확인할 수 있다. 기존의 cross entropy 손실 함수 L_{CE} 를 제외하고, 오직 본 논문에서 제안하는 L_{TS} 로만 학습시킨 proposed 3 가 가장 성능이 좋았다. 이는 교사 네트워크를 미리 학습시킬 때 L_{CE} 가 이미 사용되었고, L_{TS} 를 통해 학생 네트워크에 더 구체적인 정답을 제공해주기 때문이다.

본 논문이 제안하는 교사 네트워크의 파라미터를 초깃값으로 사용하는 방법과, L_{TS} 를 사용하여 교사-학생 학습 방법으로 학습시킨 propsed 3는 multi-style training 1 보다 clean test set에 대해서는 약 31.3% (in relative EER) 개선되었고, noisy test에 대해서도 약 17% (in relative EER) 개선되었다. proposed 3와 multi-style training 1의 DET curve 비교는 그림 5.1 과 5.2 에서 확인할 수 있다. clean test set에 대해서, proposed 3 는 multi-style training 1보다 모든 동작 포인트에서 개선된 것을 확인할 수 있다. noisy test set에 대해서는, 대략 10% False Alarm probability 동작 포인트 부근을 제외하고, 대부분의 동작 포인트에서 개선된 것을 확인할 수 있다.

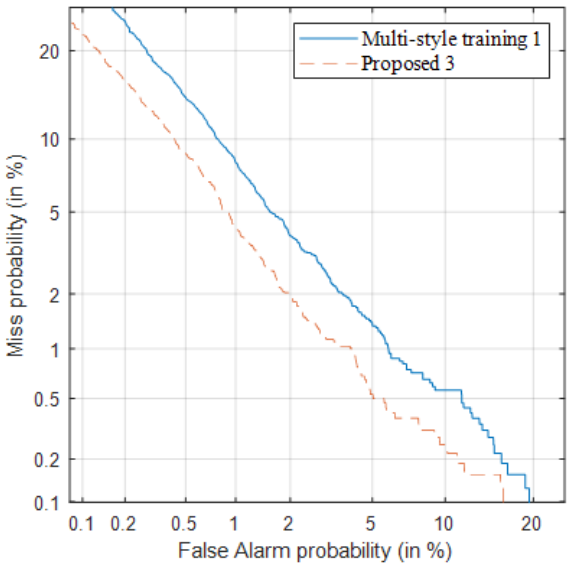

Figure 5.1: DET curves for a clean testset.

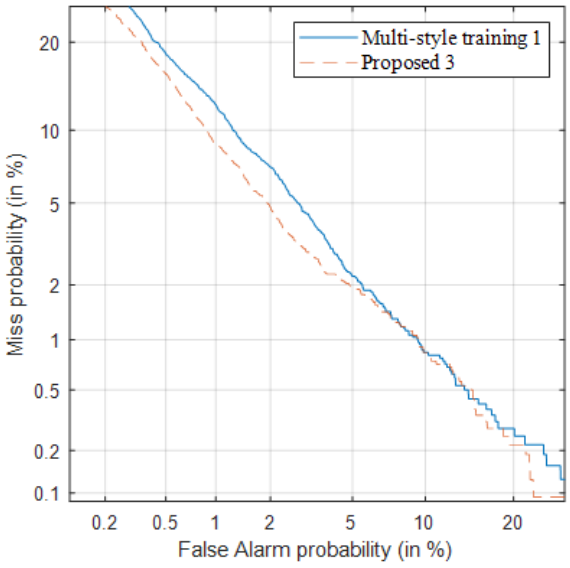

Figure 5.2: DET curves for a noisy test set.

그림 5.3 을 보면 본문이 제안하는 방법의 imitation parameter가 증가할수록 성능이 개선되는 경향을 확인할 수 있고, SNR이 증가할수록 더 많이 개선되는 것을 확인할 수 있다. multi-style 1과 비교했을 때, proposed 3은 SNR 0, 5, 10, 15, clean 환경에서 각각 3.8%, 21.6%, 27.4%, 34.3%, 31.3% (in relative EER) 개선된 것을 확인할 수 있다. 본 논문에서 제안하는 방법은 잡음에 강인함은 유지하면서도 기존의 베이스라인보다 화자 구분 능력을 개선 시킬 수 있다. 따라서 화자 임베딩이 강한 잡음에 의해 많이 손상되지 않는다면, 본 논문에서 제안하는 방법이 베이스라인보다 화자를 구분할 수 있는 정보를 더 많이 담게 된다는 것을 위의 실험 결과를 통해 알 수 있다.

앞에서 언급했던 것처럼 깨끗한 음성에 잡음을 섞어주는 방법으로 쉽게 트레이닝 데이터의 양을 늘려서 성능을 개선할 수 있다. multi-style 2 은 표 5.2 와 같이 트레이닝 데이터의 양을 두 배로 늘려서, 본 논문에서 제안하는 방법과 동일한 데이터를 사용하여 학습시켰다. 표의 multi-style 2와 multi-style 1의 결과를 보면, 이와 같은 방법을 통해 성능이 오르는 것을 확인할 수 있었지만, 동일한 데이터가 주어졌을 때 본 논문에서 제안하는 방법인 proposed 3가 더 효과적인 것을 확인할 수 있었다.

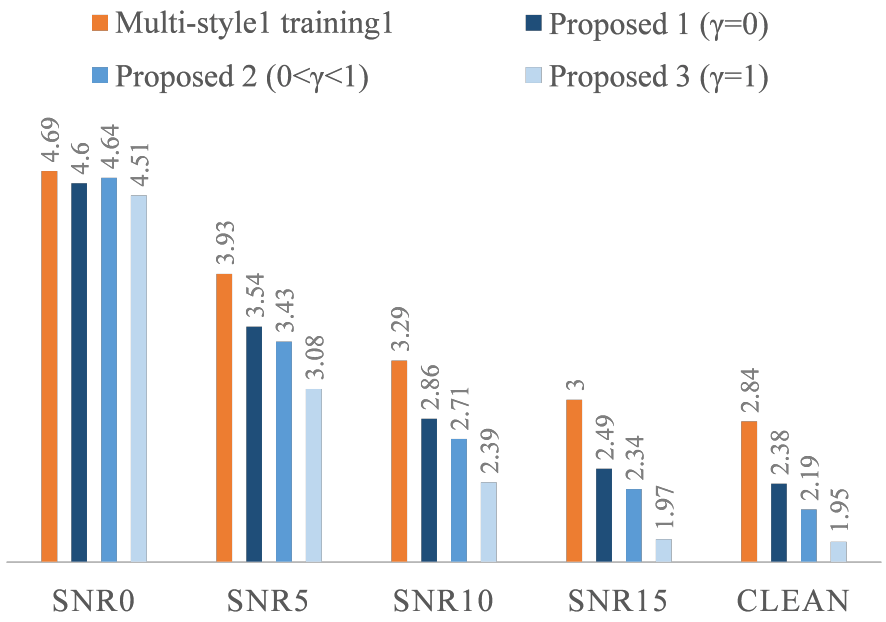

Figure 5.3: Performances under various SNR conditions. Results are reported in equal error rate (EER) [%].

Chapter 6

Conclusion and Future Work

깨끗한 음성 데이터셋과 해당 음성에 대한 잡음이 섞인 데이터셋을 사용할 수 있다면, 기존의 연구들에서는 화자 인증 시스템이 잡음에 강인해지도록 하기 위해 음성 향상 알고리즘을 전처리로 사용하거나, 잡음이 섞인 음성 데이터셋으로 네트워크를 학습시켰다. 본 논문에서는 보다 효율적인 시스템 구성을 위해 교사-학생 학습 방법을 적용한 방식을 제안하였다. 먼저 교사 네트워크를 깨끗한 음성 데이터셋으로 학습시킨 후, 잡음이 섞인 음성 데이터로 학생 네트워크를 학습시킬 때, 교사 네트워크로부터 가이드를 받을 수 있도록 구성하였다. 이 방법을 통해 학생 네트워크가 출력하는 화자 임베딩이 교사 네트워크가 출력하는 화자 임베딩과 유사해지도록 해준다. 본 논문의 실험은 휴대 기기를 사용하는 시나리오상에서 진행되었으며, 자체 녹음한 문장-종속 키워드 데이터셋과 self-attentive x-vector 모델을 사용하였다. 교사 네트워크의 파라미터를 학생 네트워크 파라미터의 초깃값으로 사용하고, 본 논문에서 제안하는 교사-학생 학습 목적 함수를 사용하여, 학생 네트워크만을 사용하여 학습시켰을 때 보다 화자 구분 능력을 개선할 수 있었고, 결과적으로 깨끗한 환경과 잡음 환경 모두에서 성능이 개선되는 것을 확인할 수 있었다.

Bibliography

[1] E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2014, pp. 4080-4084.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999-1003.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2018, pp. 5329-5333.

[4] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2018, pp. 3573-3577.

[5] 한민현, 최인규, 김남수, "임베딩 층 크기 정규화를 통한 X-vector 모델의 학습 성능 개선," 한국국방기술학회 추계학술대회, 2018, pp. 94-96.

[6] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2016, pp. 5115-5119.

[7] O. Plchot, L. Burget, H. Aronowitz, and P. Matejka, "Audio enhancing with DNN autoencoder for speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2016, pp. 5090-5094.

[8] M. Kolbœk, Z. H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," *IEEE Spoken Language Technology Workshop*, 2016, pp. 305-311.

[9] V. Vapnik and R. Izmailov, "Learning using privileged information: Similarity control and knowledge transfer," *Journal of Machine Learning Research*, vol. 16, pp. 2023-2049, 2015.

[10] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[11] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Representations*, 2015.

[12] K. Markov and T. Matsui, "Robust speech recognition using generalized distillation framework," in *Proc. Interspeech*, 2016, pp. 2364-2368.

[13] S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "Student-teacher network learning with enhanced features," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2017, pp. 5275-5279.

[14] L. Mošner, W. Minhua, A. Raju, S. H. K. Parthasarathi, K. Kumatani, S. Sundaram, R. Maas, and B. Höffmeister, "Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2019.

[15] J. H. L. Hansen, and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol.32, no.6, pp. 74-99, 2015

[16] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2016, pp. 826-834.

[17] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudeva, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265-283.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS.*, 2010.

[20] ITU, Test signals for use in telephonometry ITU-T Rec. P. 501, 2012.

[21] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, Vol.12, No.3, pp. 247-252, 1993

초 록

화자 인증 시스템은 잡음 환경에서 기본적으로 두 가지가 요구된다. 첫 번째 요구되는 특성은 화자 인증 시스템이 잡음 환경에서도 덜 손상된 화자 임베딩을 만들어내는 잡음에 강인함이다. 두 번째 요구되는 특성은 화자 인증 시스템이 만든 화자 임베딩이 화자 간의 구분이 가능한 화자 특성을 많이 포함하는 화자 구분 능력이다. 일반적으로 화자 인증 시스템이 잡음에 강인해지도록 하기 위해 모델이 학습할 트레이닝 음성 데이터에 다양한 잡음을 인위적으로 섞어준다. 이 방법은 깨끗한 음성으로만 모델을 학습시켰을 때 보다 잡음에 강인해지지만, 화자 구분 능력은 저하된다. 본 논문에서는 위에서 언급된 문제를 완화하기 위해 깨끗한 음성과 잡음이 섞인 음성 데이터를 동시에 활용하여 교사-학생 학습 모델을 구성하고 화자 인증 시스템에 적용했다. 깨끗한 음성만으로 교사 모델을 미리 학습시킨 후, 잡음이 섞인 음성으로 학생 모델을 학습시킬 때, 학생 모델이 출력하는 화자 임베딩과 교사 모델이 출력하는 화자 임베딩이 같아지게끔 하는 손실 함수를 추가하여 교사 모델로부터 정보를 얻을 수 있도록 하였다. 본 논문의 실험은 휴대 기기를 사용하는 시나리오상에서 진행되었으며, 자체적으로 수집한 키워드 데이터셋과 self-attentive x-vector를 사용하였다. 실험 결과를 통해 교사-학생 학습 방법이 화자 구분 능력이 저하되는 것을 완화할 수 있었고, 결과적으로 잡음환경과 깨끗한 환경 모두에서 성능이 오르는 것을 확인하였다.

주요어: 잡음에 강인한 화자 인증 시스템, 교사-학생 학습 방법
학 번: 2017-22872